

# Herramientas para el diseño y el análisis de datos de paneles de genes

**Hospital Sant Pau**  
**Barcelona, 16 Jun 2016**



PRINCIPE FELIPE  
CENTRO DE INVESTIGACION

Computational • Genomics



## Goal: biomedical research

- **Basic research** in genes, targets, molecular and cellular processes, Nanomedicine and Computational Medicine
- **Translation into clinical practice:** personalized medicine, cancer, rare diseases, metabolic and functional impairment

*<http://www.cipf.es/>*

# Who are we?

---

- The **Computational Genomics** Department, in Research Center Prince Felipe
- **Team:** multidisciplinary group of 14 researchers and technicians led by Joaquín Dopazo

*<http://bioinfo.cipf.es/>*

# Who are we?



**Introduction**

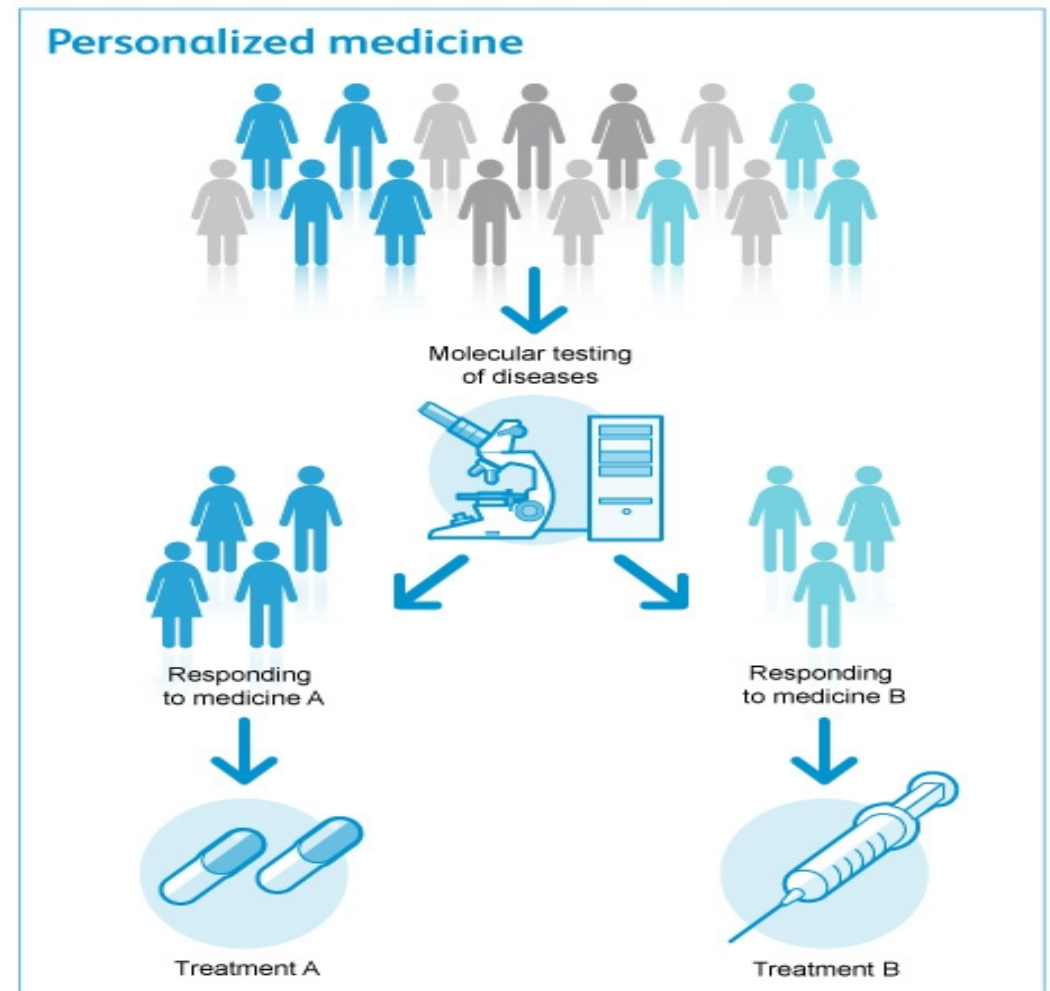
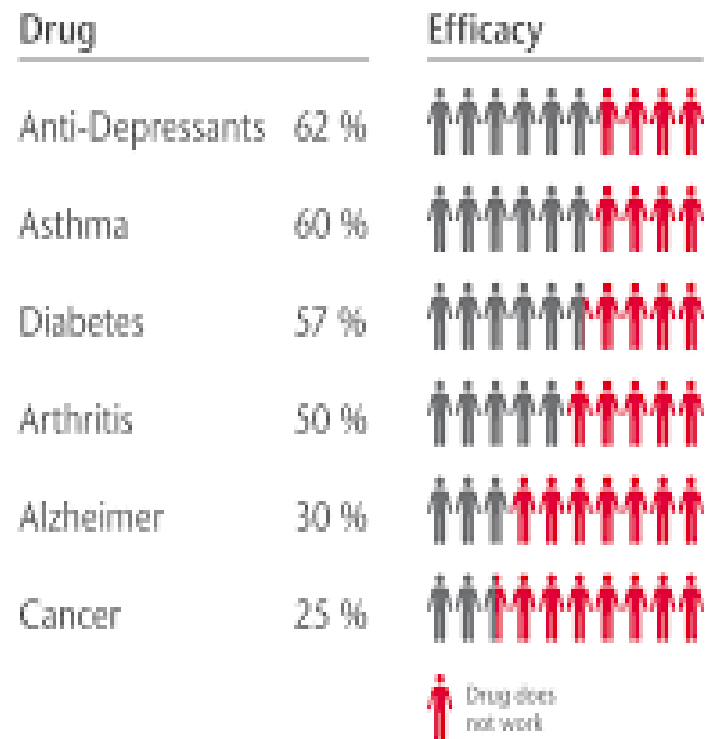
**Genomic Computational Department**

# Why are we interested in Computational Genomics?

---

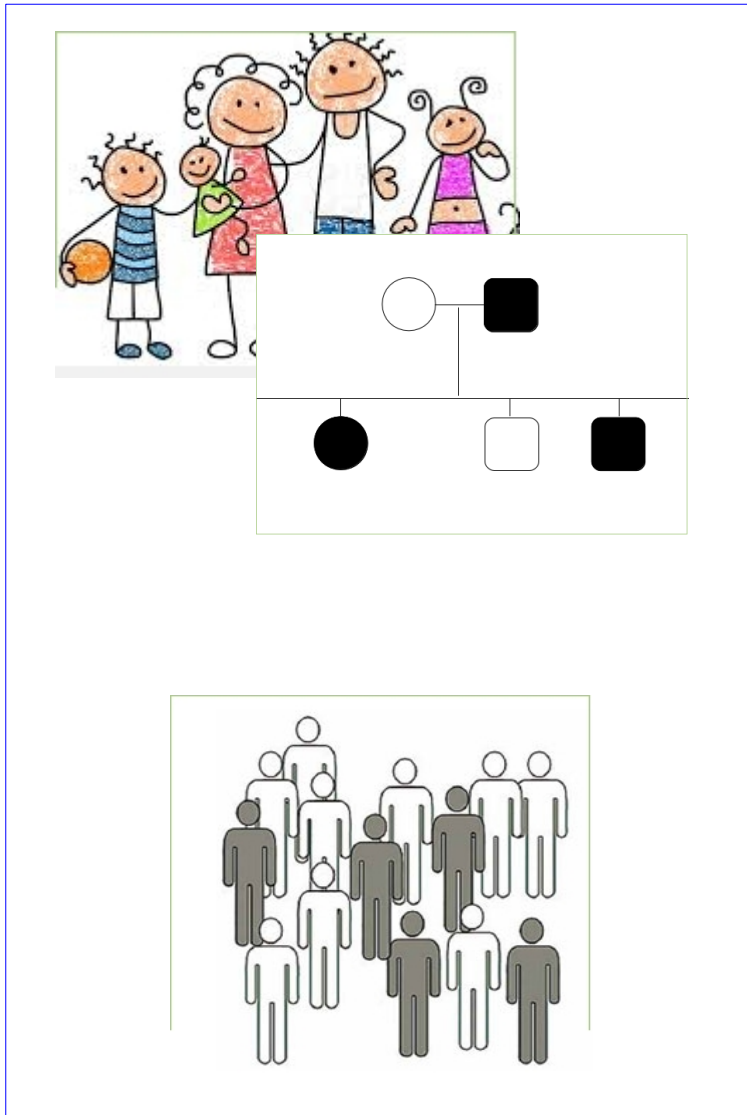
- The **overall goal** of the department:
  - Apply computational methods to biomedical and biotechnological problems
- **Research interests:**
  - The development and application of novel bioinformatics **methods** aimed at **discovering new drugs**
  - Identification of genes or proteins may be considered **therapeutic targets**
  - **Personalized medicine:** tools for discovering and diagnostic

# Why are we interested in Computational Genomics?



New molecular and diagnostic technologies can be used to match select groups of patients with treatments that may give them the best results

# Why are we interested in Computational Genomics?

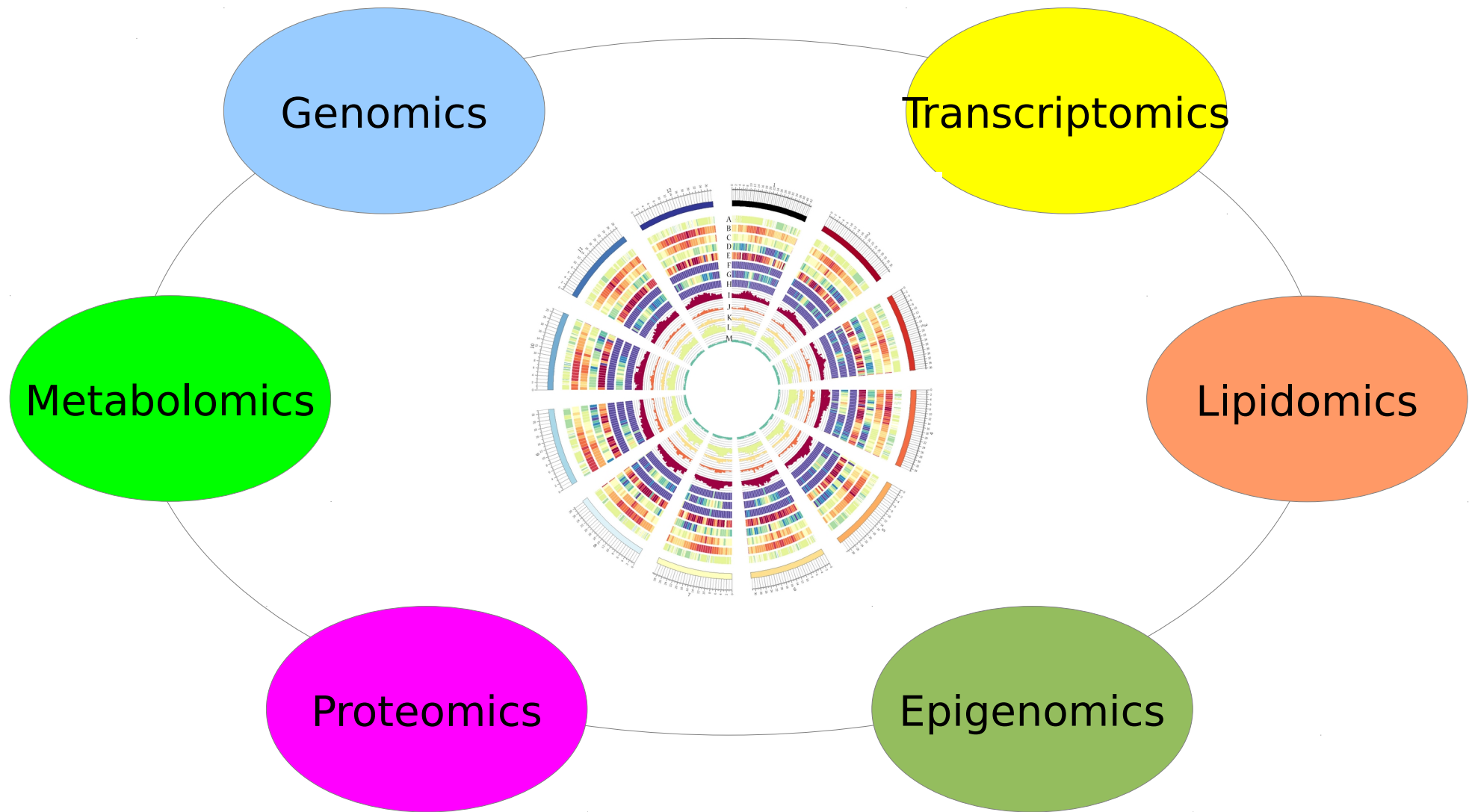


Introduction

Personalized Medicine and Mendelian Diseases



# Big Data

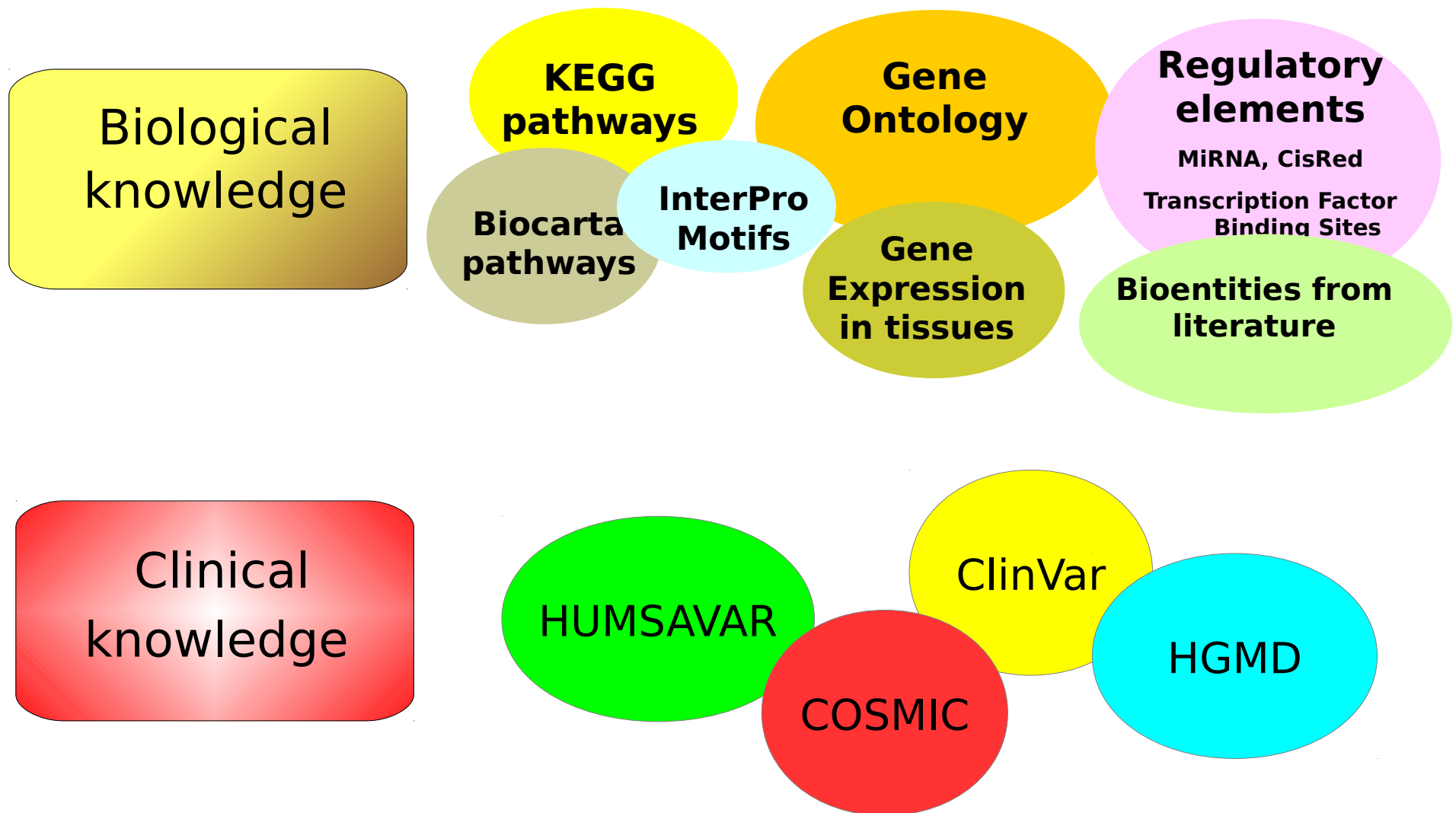


**Introduction**

**Omics sciences**



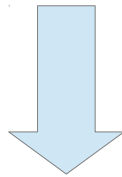
# Big Data



# How do we work?

- Our department collaborates in different research projects and converts researcher needs into bioinformatics solutions
- Free software for several reasons:
  - **Any customer can try our tools**
  - **The scientific community can test our software**
  - **This is the current trend in Computational Genomics**

# How do we work?



Introduction

Genomic Computational Department

# How do we work?

- El CIBER en su Área Temática de Enfermedades Raras (CIBERER) es el **centro de referencia** en España en investigación sobre **enfermedades raras**: <http://www.ciberer.es/>
- **Objetivo**: coordinar y favorecer la investigación básica, clínica y epidemiológica, así como potenciar que la investigación que se desarrolla en los laboratorios llegue al paciente, y dé respuestas científicas a las preguntas nacidas de la interacción entre médicos y enfermos.
- El CIBERER se compone de un equipo humano de más de 700 profesionales e integra a **62 grupos de investigación**.



# How do we work?

---

- Curso CIBERER de análisis de datos genómicos, **28-30 Sep 2016** en Valencia:  
<http://bioinfo.cipf.es/mda15ciberer>
- International course of Genomic Data Analysis, **Mar 2017**, Valencia: <http://bioinfo.cipf.es/gda16/program/>
- <http://bioinfo.cipf.es/courses>

# Web tools to analyze gene panel data



Computational • Genomics



# Outline

---

## 1) Introduction to NGS Data Analysis

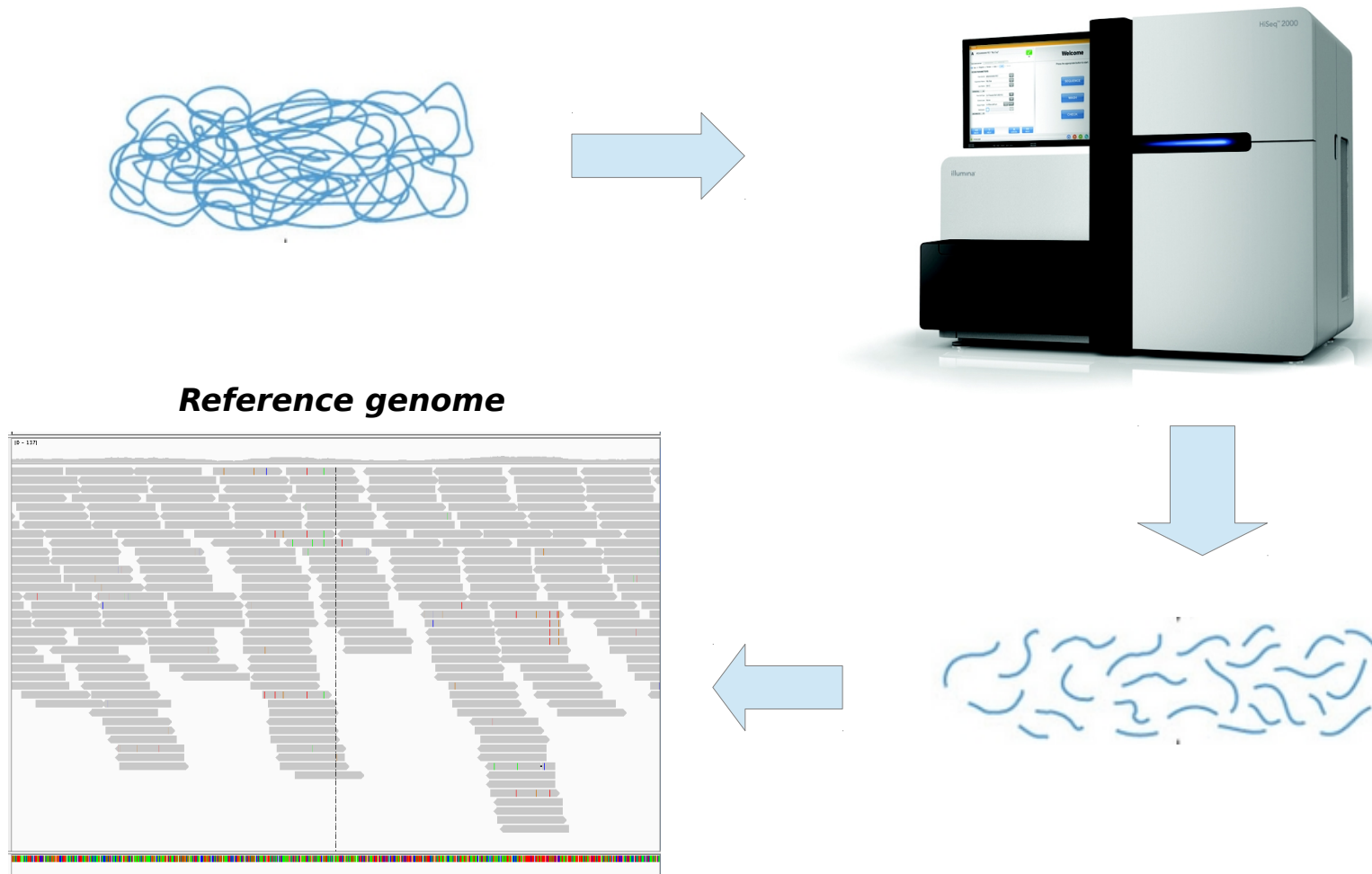
## 2) TEAM

## 3) PanelMaps

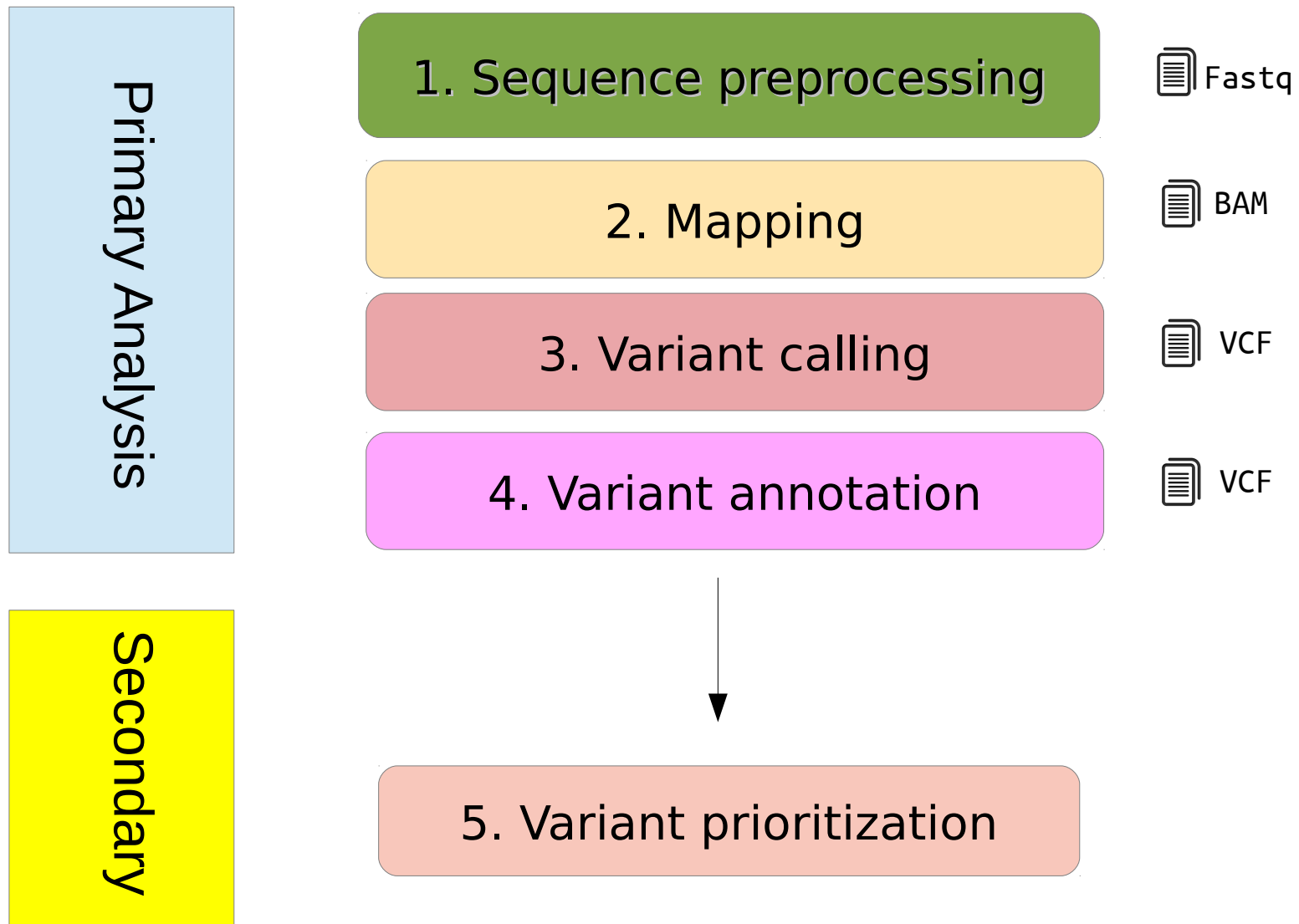


# NGS technologies

How do these technologies work ?



# Genomics Data Analysis Pipeline



# Fastq format

- We could say “it is a fasta with **qualities**”:
  - 1. Header (like the fasta but starting with “@”)
  - 2. Sequence (string of nt)
  - 3. “+” and sequence ID (optional)
  - 4. Encoded quality of the sequence

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%+))((%%)).1***-+*''))*55CCF>>>>>CCCCCCC65
```

# BAM/SAM format

```
@PG ID:HPG-Aligner VN:1.0
@SQ SN:20 LN:63025520

HWI-ST700660_138:2:2105:7292:79900#2@0/1 16 20 76703 254 76= * 0 0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
GIJGJLGGFLILGGIEIFEKEDELIGLJIHJFIKKFELFIKLFFGLGHKKGJLFIIGKFFEFFEFGKCKFHHCCCF AS:i:254 NH:i:1 NM:i:0

HWI-ST700660_138:2:2208:6911:12246#2@0/1 16 20 76703 254 76= * 0 0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
HHJFHLGFFLILEGIKIEEMGEDLIGLHIHJFIKKFELFIKLEFGKGHEKHJLFHIGKFFDFFEF GKDKFHHCCCF AS:i:254 NH:i:1 NM:i:0

HWI-ST700660_138:2:1201:2973:62218#2@0/1 0 20 76655 254 76M * 0 0
AACCCCAAAAATGTTGGAAGAATAATGTAGGACATTGCAGAAGACGATGTTTAGATACTGAAAGGGACATACTTCT
FEFFGHHHGGHFKCCJKFHIGIFFIFLDEJKGJGGFKIHLFIJGIEGFLDEDFLFG EIIMHHIKL$BBGFFJIEHE AS:i:254 NH:i:1 NM:i:1

HWI-ST700660_138:2:1203:21395:164917#2@0/1 256 20 68253 254 4M1D72M * 0 0
NCACCCATGATAGACCAGTAAAGGTGACCACTTAAATTCCTTGCTGTGCAGTGTTCTGTATTCTCAGGACACAGA
#4@ADEHFJFFEJDHJGKEFIHGHGBGFHHFIICEIIFFKIFHEGJEHHGLELEGKJMFGGGLEIKHLFGKIKHDG AS:i:254 NH:i:3 NM:i:1

HWI-ST700660_138:2:1105:16101:50526#6@0/1 16 20 126103 246 53M4D23M * 0 0
AAGAAGTGCAAACCTGAAGAGATGCATGTAAAGAATGGTTGGGCAATGTGCGGCAAAGGGACTGCTGTGTTCCAGC
FEHIGGHIGIGJI6FCFHJIFFLJJCJGJHGFKKKKGJIKHFFKIFFFKHFLKHGKJLJGKILLEFFLIHJIEIB AS:i:368 NH:i:1 NM:i:4
```

## SAM Specification:

<http://samtools.sourceforge.net/SAM1.pdf>

Introduction

NGS data analysis: files format

# VCF format

```
#fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

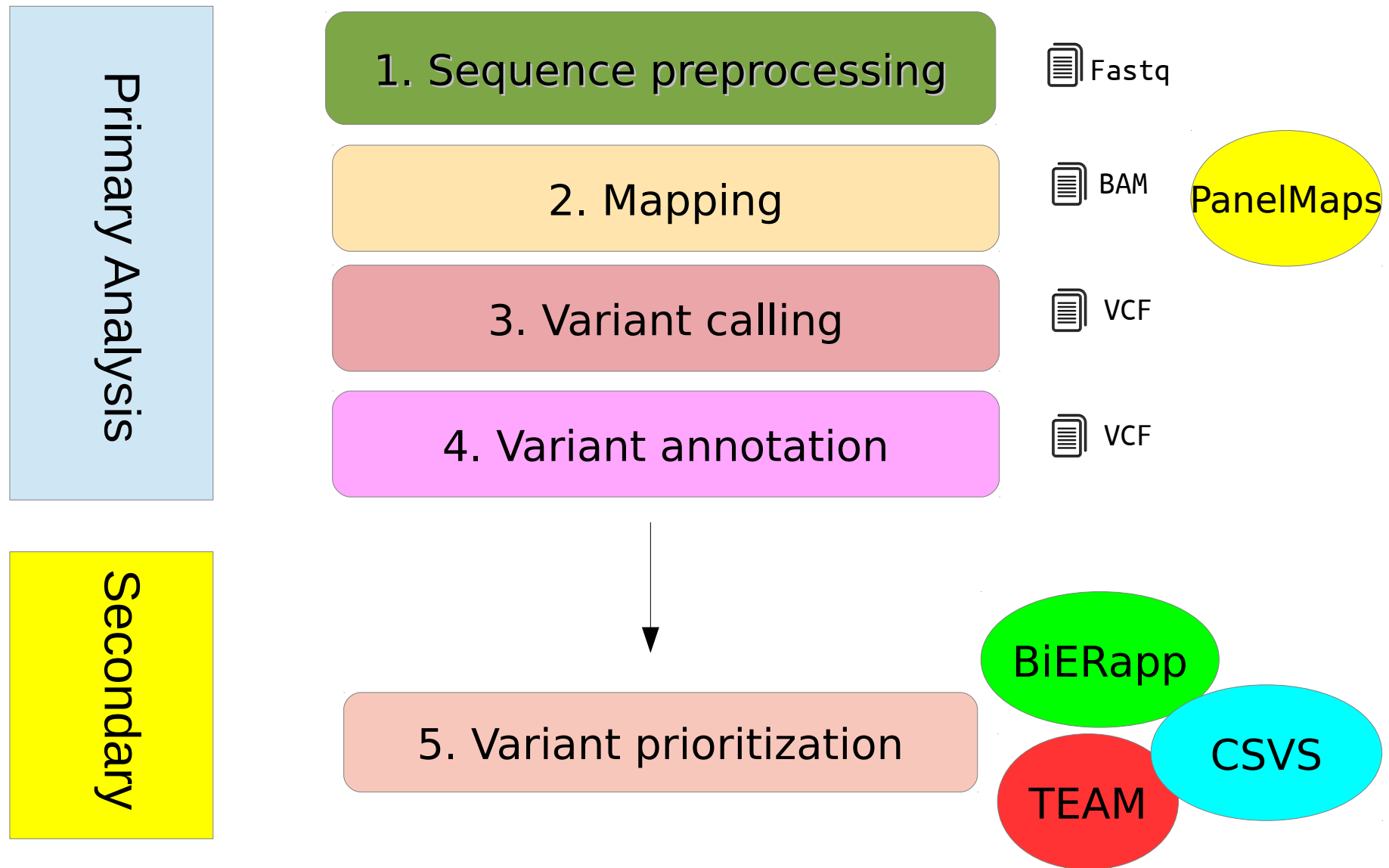
<http://www.1000genomes.org/>

# BED format

|      |           |           |      |   |   |           |           |         |
|------|-----------|-----------|------|---|---|-----------|-----------|---------|
| chr7 | 127471196 | 127472363 | Pos1 | 0 | + | 127471196 | 127472363 | 255,0,0 |
| chr7 | 127472363 | 127473530 | Pos2 | 0 | + | 127472363 | 127473530 | 255,0,0 |
| chr7 | 127473530 | 127474697 | Pos3 | 0 | + | 127473530 | 127474697 | 255,0,0 |
| chr7 | 127474697 | 127475864 | Pos4 | 0 | + | 127474697 | 127475864 | 255,0,0 |
| chr7 | 127475864 | 127477031 | Neg1 | 0 | - | 127475864 | 127477031 | 0,0,255 |
| chr7 | 127477031 | 127478198 | Neg2 | 0 | - | 127477031 | 127478198 | 0,0,255 |
| chr7 | 127478198 | 127479365 | Neg3 | 0 | - | 127478198 | 127479365 | 0,0,255 |
| chr7 | 127479365 | 127480532 | Pos5 | 0 | + | 127479365 | 127480532 | 255,0,0 |
| chr7 | 127480532 | 127481699 | Neg4 | 0 | - | 127480532 | 127481699 | 0,0,255 |

<https://genome.ucsc.edu/FAQ/FAQformat.html#format1>

# Genomics Data Analysis Pipeline





# Outline

---

1) Introduction to NGS Data Analysis

2) **TEAM**

3) PanelMaps

# Can I interpret sequencing data for diagnostic?

<http://team.babelomics.org/beta/>



Computational • Genomics



TEAM

Targeted Enrichment Analysis and Management

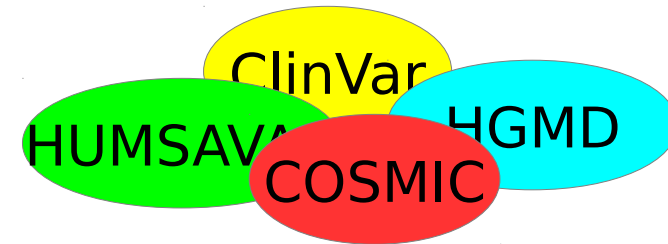
# Introduction

Sequencing  
data

Biological  
knowledge



TEAM

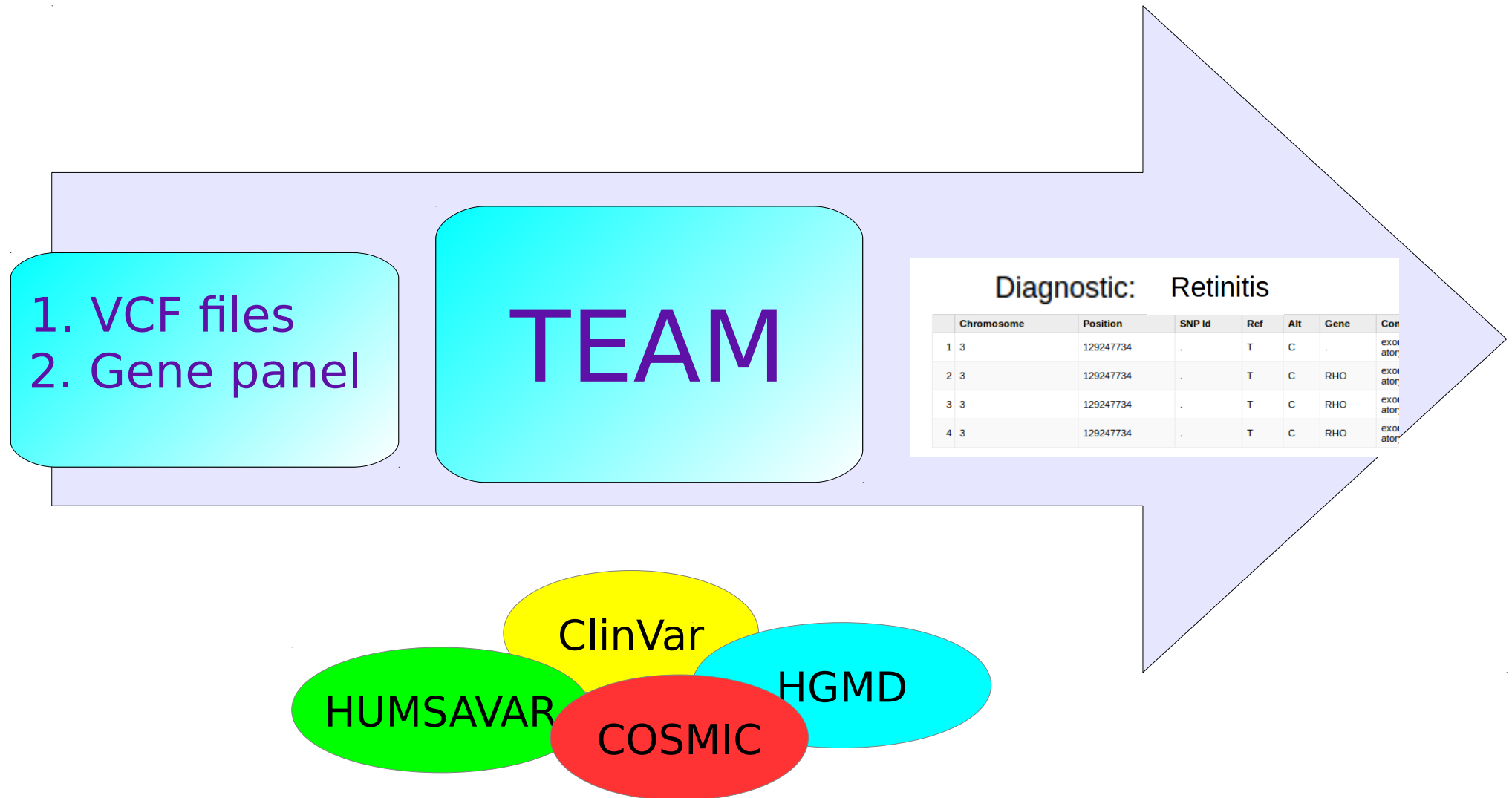


Diagnostic

TEAM

Targeted Enrichment Analysis and Management

# How does TEAM work?



**TEAM**

**Targeted Enrichment Analysis and Management**

# Getting information

## □ SIFT

- SIFT predicts whether an amino acid substitution affects protein function
- **Interpretation:** 1 (tolerated) to 0 (not tolerated)

<http://sift.jcvi.org/>



## □ PolyPhen

- Polymorphism Phenotyping is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein
- **Interpretation:** 1 (probably damage) to 0 (benign)

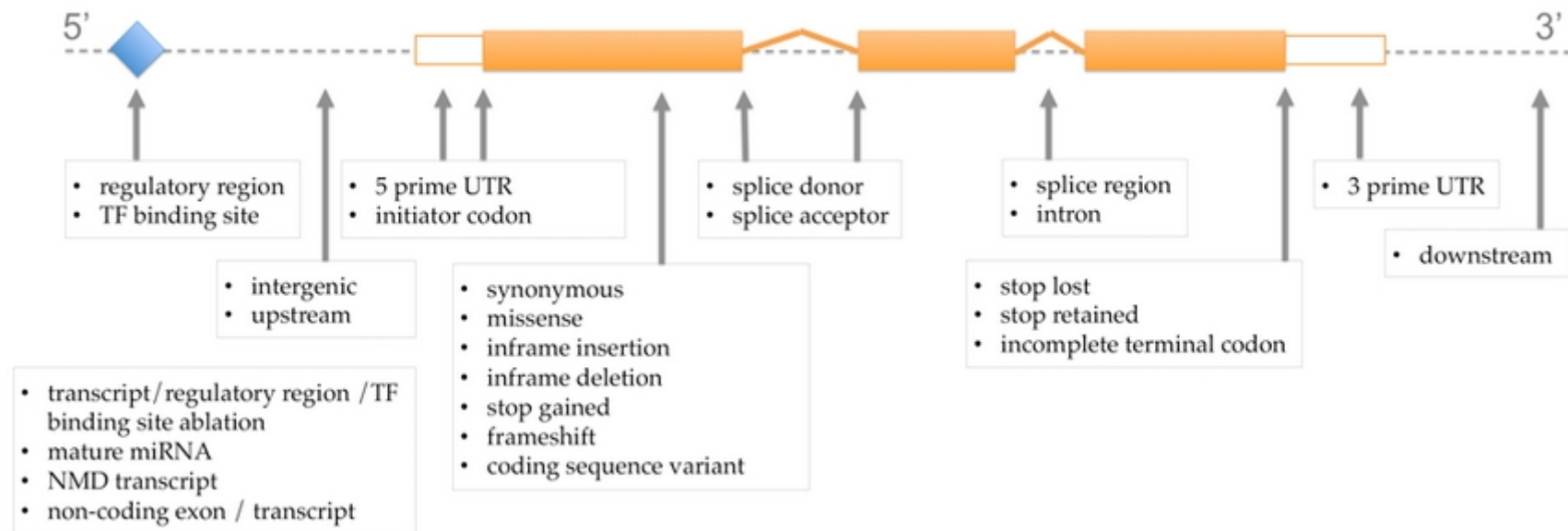
<http://genetics.bwh.harvard.edu/pph2/index.shtml>



# Getting information

The screenshot shows the Ensembl website interface. At the top, there's a navigation bar with links: BLAST/BLAT, BioMart, Tools, Downloads, and Help & Documentation. Below this, a secondary navigation bar includes 'Using this website', 'Annotation & prediction' (which is highlighted), 'Data access', 'API & software', and 'About us'. On the left, a sidebar titled 'In this section' lists: Data Description, Predicted Data, Import VCF script, and Variation Sources. The main content area is titled 'Ensembl Variation - Predicted data' and shows a breadcrumb trail: Home > Help & Documentation > Annotation & Prediction.

## Consequence type or effect



[http://www.ensembl.org/info/genome/variation/predicted\\_data.html](http://www.ensembl.org/info/genome/variation/predicted_data.html)

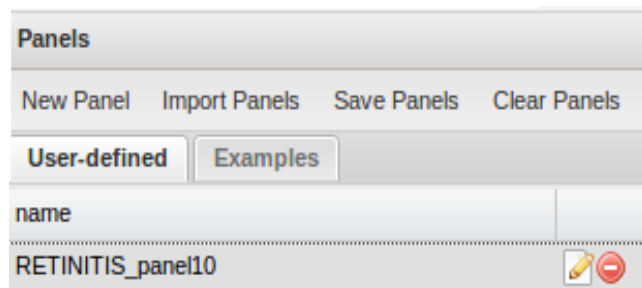
TEAM

Targeted Enrichment Analysis and Management

# How does TEAM work?

<http://team.babelomics.org/beta/>

## 1. Defining panel



Panels

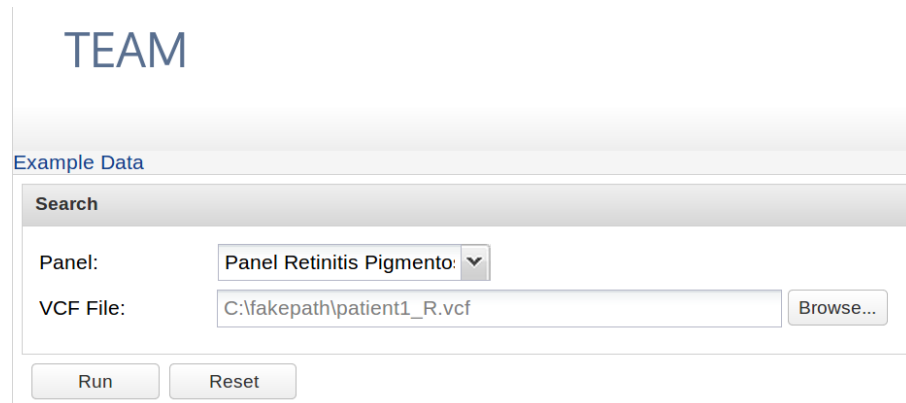
New Panel Import Panels Save Panels Clear Panels

User-defined Examples

name

RETINITIS\_panel10

## 2. Uploading input data



TEAM

Example Data

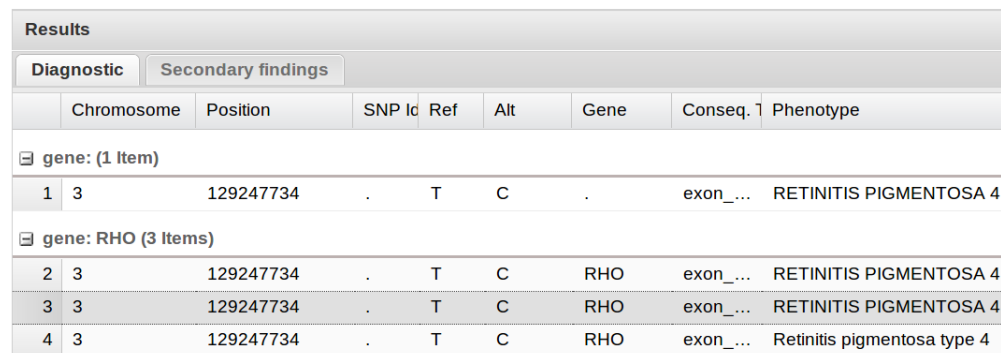
Search

Panel: Panel Retinitis Pigmento

VCF File: C:\fakepath\patient1\_R.vcf Browse...

Run Reset

## 3. Getting results



| Results                       |            |           |        |     |     |      |                                      |
|-------------------------------|------------|-----------|--------|-----|-----|------|--------------------------------------|
| Diagnostic Secondary findings |            |           |        |     |     |      |                                      |
|                               | Chromosome | Position  | SNP Id | Ref | Alt | Gene | Conseq. 1 Phenotype                  |
| gene: (1 item)                |            |           |        |     |     |      |                                      |
| 1                             | 3          | 129247734 | .      | T   | C   | .    | exon_... RETINITIS PIGMENTOSA 4      |
| gene: RHO (3 items)           |            |           |        |     |     |      |                                      |
| 2                             | 3          | 129247734 | .      | T   | C   | RHO  | exon_... RETINITIS PIGMENTOSA 4      |
| 3                             | 3          | 129247734 | .      | T   | C   | RHO  | exon_... RETINITIS PIGMENTOSA 4      |
| 4                             | 3          | 129247734 | .      | T   | C   | RHO  | exon_... Retinitis pigmentosa type 4 |

TEAM

Targeted Enrichment Analysis and Management



# How to define a panel?

1. Name of panel

2. Diseases

3. Adding:  
- more genes  
- mutations

4. Save panel

The screenshot shows the 'Panel Manager' window with the following components:

- Name:** A text field containing 'RETINITIS\_10'.
- Diseases (Drag):** A list of diseases including 'Ataxia\_and\_retinitis\_pigmentosa\_with\_isolate...', 'Hypoprebetalipoproteinemia,\_acanthocytosis,...', 'Juvenile\_retinitis\_pigmentosa,\_AIPL1-related', 'Neuropathy\_ataxia\_retinitis\_pigmentosa\_syn...', 'POSTERIOR COLUMN ATAXIA WITH RETINI...', 'Polyneuropathy, hearing loss, ataxia, retinitis ...', and several 'RETINITIS PIGMENTOSA' variants (1, 11, 12, 14, 17, 18, 19, 2, 25, 26, 27). An arrow points to 'RETINITIS PIGMENTOSA 14'.
- Primary Disease (Drop):** A list containing 'RETINITIS PIGMENTOSA 10', 'RETINITIS PIGMENTOSA 13', and 'RETINITIS PIGMENTOSA 20'.
- Genes:** A list containing 'IMPDH1', 'PRPF8', and 'RPE65', each with a red minus button to its right.
- Mutations:** A table with columns 'Chr', 'Pos', 'Ref', 'Alt', and 'Gene'. An arrow points to the table area.
- Text/Bed File:** Radio buttons for 'Text' (selected) and 'Bed File'. Below is a text field containing 'BRCA2,PPL'. An arrow points to this field.
- Buttons:** 'Add Mutation' and 'Add Genes' buttons are located below the mutations and genes lists respectively. An arrow points to the 'Add Genes' button.
- Footer:** 'PolyPhen:' and 'Sift:' dropdown menus, and 'Add new panel', 'Clear', and 'Close' buttons.

# How to define a panel?

The screenshot shows a web-based interface for adding mutations. At the top, there is a form with fields for 'Chr:' (8), 'Pos:' (55539395), 'Ref:' (A), 'Alt:' (T), 'Gene Name:' (RP1), and 'Disease Name:' (Lung cancer 2). Below these fields are buttons for 'Reset', 'Check', and 'Add Mutation'. A blue arrow points from the text 'Adding new mutations' to the 'Add Mutation' button. Below the form is a 'Region overview' section with a 'Window size: 583 nts'. It displays a genomic track with coordinates 55,539,104 to 55,539,686. A zoomed-in view shows the sequence 'AA G C A C A T A A C T A A A A T T G C C G G T T T G A C A G G A G A T A A T C T A T G T A A A G A G G G A G A T A A G T C T T T' with a vertical orange line at position 55,539,395. A blue arrow points from the text 'Checking mutations from Genome Viewer' to this zoomed-in view. Below the sequence is a 'Gene' track showing two red lines. At the bottom is an 'SNP' track with various markers like P\_ESP\_8\_55539357, rs58051614, rs200135800, COSM486527, and others. A status bar at the bottom right indicates 'T 8:55,539,394 Genome Viewer'.

Adding  
new mutations

Checking  
mutations from  
Genome Viewer

TEAM

Targeted Enrichment Analysis and Management

# Web results

## 1. OVERVIEW

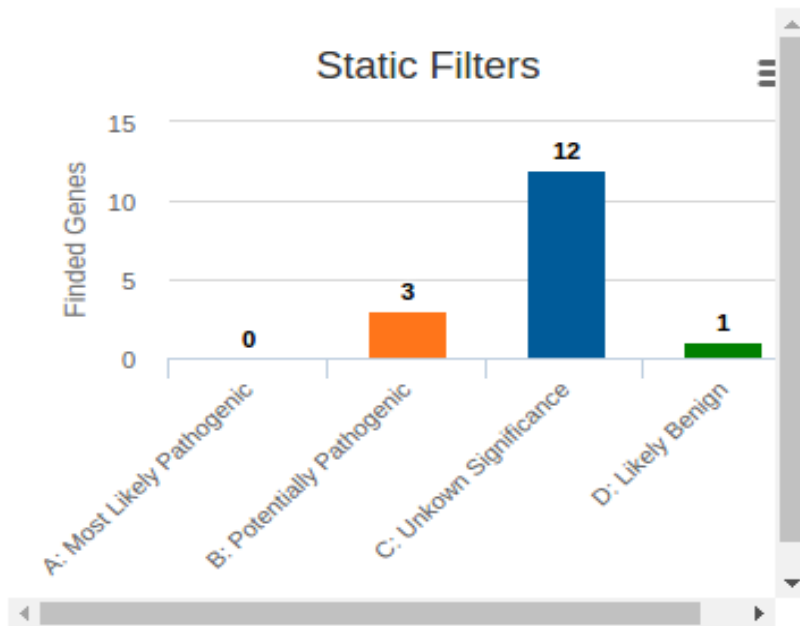
prueba\_charcot  
New Diagnosis

Total Variants: 17

Diagnostic Variants: 1

Secondary Variants: 16

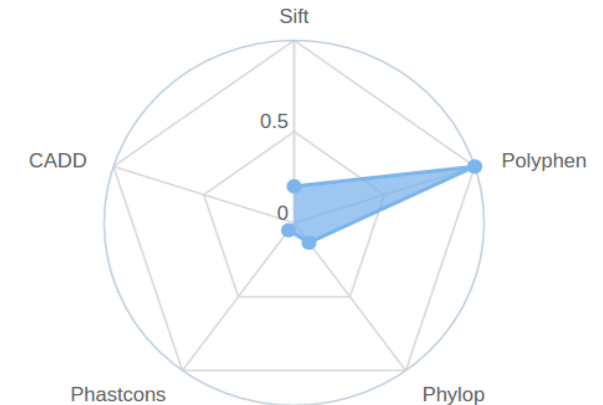
Static Filters



## 2. DIAGNOSTIC

Diagnostic

| Chr | Pos      | Ref | Alt | Gt | SNP Id |
|-----|----------|-----|-----|----|--------|
| X   | 70444245 | C   | T   | .  | GJB1   |



TEAM

Targeted Enrichment Analysis and Management

# Web results

## 3. SECONDARY FINDINGS

### Variant Filters

Static Filters

Custom Filters

Without Filters

▼ 16

A: Most Likely Pathogenic

▼ 0

B: Potentially Pathogenic

▼ 3

C: Unknown Significance

▼ 12

D: Likely Benign

▼ 1

### Secondary Findings

| Chr | Pos       | Ref | Alt | Gt | SNP Id     |
|-----|-----------|-----|-----|----|------------|
| 1   | 10435324  | C   | A   | .  | KIF1B      |
| 1   | 156107534 | C   | T   | .  | LMNA       |
| 1   | 10318652  | C   | G   | .  | KIF1B      |
| 5   | 148386525 | T   | C   | .  | SH3TC2     |
| 14  | 102454933 | C   | A   | .  | DYNC1H..   |
| 14  | 102514227 | T   | C   | .  | DYNC1H..   |
| 5   | 148407708 | A   | C   | .  | SH3TC2     |
| 5   | 148408101 | A   | G   | .  | SH3TC2     |
| 1   | 10355834  | C   | T   | .  | KIF1B,R... |
| 14  | 102515015 | G   | A   | .  | DYNC1H..   |

Static Filters

Custom Filters



Clear

Search

Position

Chromosomal location:

1:1-1000000,2:1-1000000

Gene:

BRCA2, PPL

SNPId:

rs998817

Population Freqs. +

Genotype +

Quality +

Protein Substitution Scores +

Conservation +

Consequence Type +

TEAM

Targeted Enrichment Analysis and Management

# Reporting results

## 4. REPORT PDF

### Select to Show

- ☒ Generic Data
- ☒ Static Filter Resume Table
- ☒ Custom Filter Resume Table
- ☒ Additional Patient Data
- ☐ Editable Conslusions
- ☒ Diagnostic Table
- ☐ Most Pathogenic Table
- ☐ Secondary Findinds (with last custom filter used)
- ☒ Annex I: Static Filters (decision umbrals)
- ☐ Annex II: Panels



prueba\_charcot

### New Diagnosis

Patient Diagnostic:

|                     |    |
|---------------------|----|
| Total Variants      | 17 |
| Diagnostic Variants | 1  |
| Secondary Variants  | 16 |

| Static Filters              | Variants Found |
|-----------------------------|----------------|
| ⚠ A: Most Likely Pathogenic | 0              |
| ⚠ B: Potentially Pathogenic | 3              |
| ⚠ C: Unkown Significance    | 12             |
| ✅ D: Likely Benign          | 1              |

| Custom Filter Used | Variants Found |
|--------------------|----------------|
| -No filter used-   | 16             |

TEAM

Targeted Enrichment Analysis and Management

# Remarks

---

- TEAM is a **free** web tool
- **Easy-to-use** and **powerful**
- TEAM helps you for **diagnostic**

# More information

Nucleic Acids Research Advance Access published May 26, 2014

*Nucleic Acids Research*, 2014 **1**  
doi: 10.1093/nar/gku472

## A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications

Alejandro Alemán<sup>1,2</sup>, Francisco Garcia-Garcia<sup>1</sup>, Ignacio Medina<sup>1</sup> and Joaquín Dopazo<sup>1,2,3,\*</sup>

<sup>1</sup>Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, 46012, Spain,

<sup>2</sup>Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, 46012, Spain and

<sup>3</sup>Functional Genomics Node, (INB) at CIPF, Valencia, 46012, Spain



TEAM Tutorial:

<http://ciberer.es/bier/team>



TEAM

Targeted Enrichment Analysis and Management



# Outline

---

1) Introduction to NGS Data Analysis

2) TEAM

**3) PanelMaps**

# Can I visualize and detect deletions for gene panel?



Computational • Genomics



PanelMaps

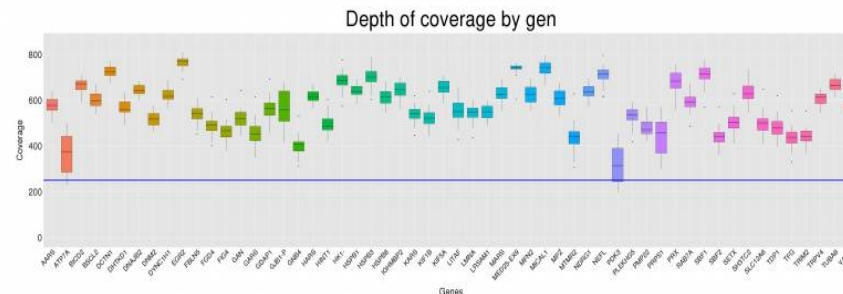
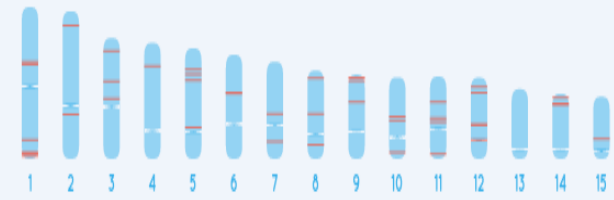
A web tool to analyze gene panel data

# How does PanelMaps work?

1. BAM files  
2. BED file

PanelMaps

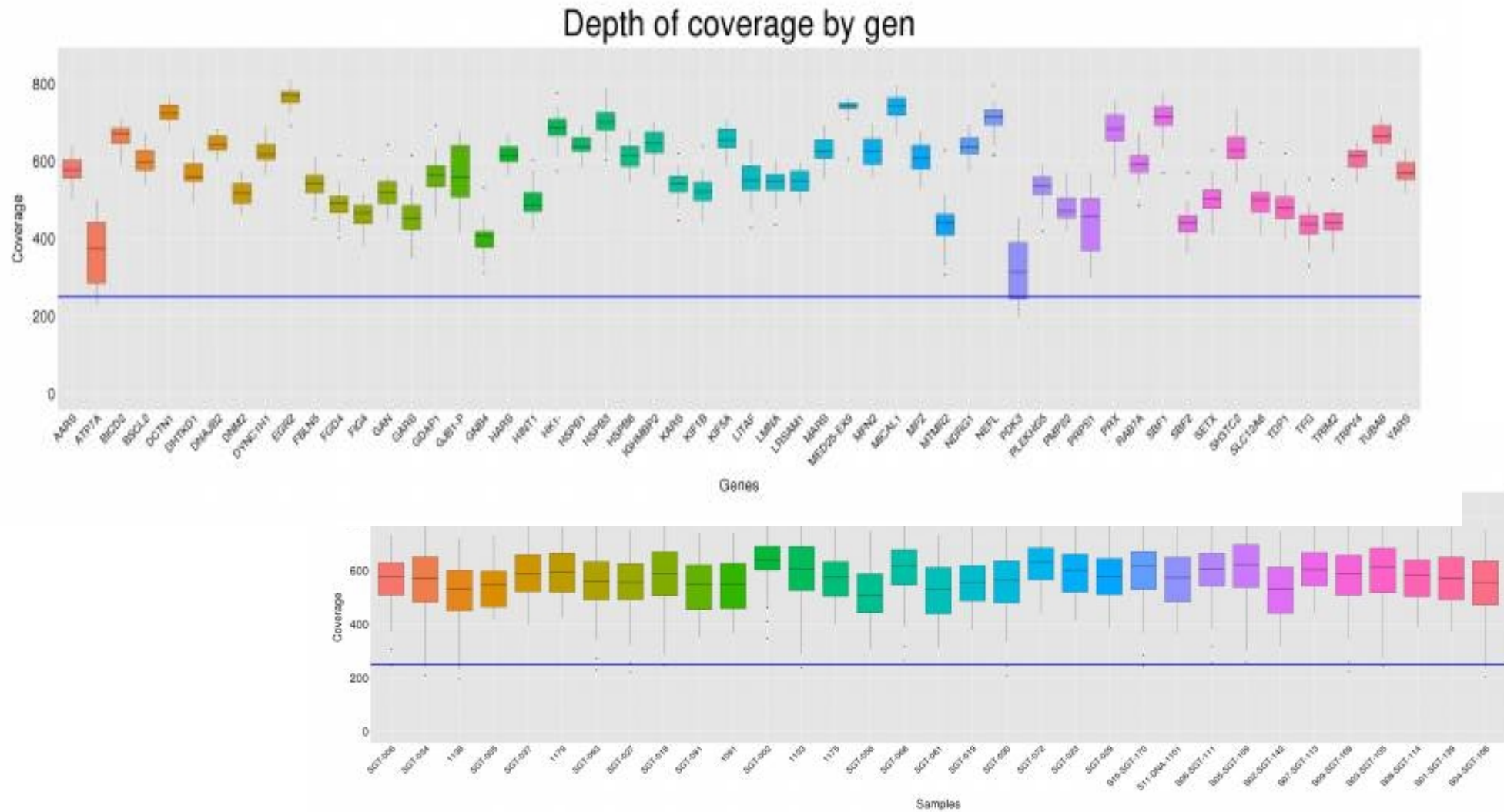
Karyotype *hsapiens* (grch37)



PanelMaps

Detecting altered regions for targeted sequencing

# How does PanelMaps work?



# PanelMaps

## Coverage description

# How does PanelMaps work?

PanelMaps

Worked example

Download

Docs

## Chromosome

Select a chromosome where you want to find regions:

All chromosomes ▼

## Select control

Select the control sample (optional):

No control ▼

## Threshold

2

## Min window size

10

## Min cover value

200

Find regions

Title: Test 9 samples

Description: Test PanelMaps with 9 samples



Region name



Labels



Karyotype hsapiens (grch37)



Coverage

Use the upper panel to select a chromosome. Then, select the region of interest.



Genes

<http://panelmaps.juanes.xyz/dashboard/PMAPSDemo>

PanelMaps

Detection of altered regions

# How does PanelMaps work?



<http://panelmaps.juanes.xyz/dashboard/PMAPSDEMO>

**PanelMaps**

**Detecting altered regions for targeted sequencing**

# Any comment or question?

---



**Web tools for gene panel data**